Does Road Condition Affect the Severity of Injury in Bicycle Accidents?

Causal Inference course

Maayan Aytek and Omri Itzhaki

Faculty of Data and Decision Sciences

TECHNION – ISRAEL INSTITUTE OF TECHNOLOGY

October 2024

GitHub Repository:
BikeCrashCausalInference on GitHub

Contents

1 Introduction of the Problem

Bicycling serves as a vital mode of transportation, physical activity, and recreation worldwide. The increasing popularity of cycling, especially in urban areas, contributes to benefits such as reduced air pollution, decreased traffic congestion, and improved public health outcomes. Despite these advantages, cyclists are among the most vulnerable road users, with a significantly higher risk of injury and fatality when involved in accidents compared to occupants of motor vehicles. According to the National Security Council (NSC), the number of preventable deaths from bicycle-related incidents increased by 10% in 2022 and has risen 47% over the last decade.

Studies indicate that road surface conditions, specifically whether the surface is wet or dry, can significantly affect accident outcomes. For instance, research by Asgarzadeh, Morteza, et al. (2018) found that the risk of severe injuries in bicycle accidents increased by 10% during rainy conditions, with slippery road surfaces and reduced braking effectiveness cited as primary contributors. Similarly, studies like Nyberg, Peter, Ulf Björnstig, and L. O. Bygren (1996) and Janstrup, Kira Hyldekær, Mette Møller, and Ninette Pilegaard (2018) have shown that wet or damaged road conditions are associated with a higher incidence of serious and fatal injuries for cyclists, often due to diminished tire grip and increased stopping distances.

By examining the effect of road conditions on the severity of cyclist injuries, this study aims to contribute to a better understanding of potential risk factors associated with adverse weather. Specifically, we will try to estimate the effect of road condition (T) on injury severity (Y). The findings may help inform improvements in road safety measures, such as encouraging cyclists to exercise added caution on wet roads and considering minor infrastructural adjustments to enhance traction in high-risk areas. Additionally, this research could support local efforts to raise awareness about safe cycling practices, particularly in wet weather, thereby helping to reduce the likelihood of severe injuries among cyclists.

2 The Data

2.1 Description

The dataset was sourced from the Open Data platform for the town of Chapel Hill and is accessible on Kaggle. It contains detailed records of 11,266 bicycle crashes in North Carolina from 2007 to 2018. The dataset comprises 60 attributes that capture critical details for each incident, including the time, location, severity, and nature of the crash. Additionally, it provides information on the bicyclists involved, such as age, gender, injury severity, alcohol or drug involvement, and travel direction, along with details on the vehicles' drivers, including demographic information, vehicle type, and substance use. Environmental and road conditions are also documented, encompassing factors such as road type and surface, lighting, weather, speed limits, and the presence of traffic control measures at the time of each incident.

2.2 Features and Pre-processing

The dataset contained numerous features, some of which were irrelevant to the task, while others included a significant number of missing or 'unknown' values. We excluded features where missing values were predominant and pre-processed the remaining feature values to enhance usability. Rows with excessive missing values were also removed. Additionally, we consolidated certain feature values that were overly detailed for our purposes to reduce unnecessary noise. After these procedures, a total of 8,154 records remained, and the following feature list was retained:

- **Age category**: Represents the age range of the biker, with values spanning 0-5, 6-10, 11-15, 16-19, 20-24, 25-29, 30-39, 40-49, 50-59, 60-69, and 70+. We transformed this feature into a numeric format by assigning each interval its lower bound.
- **Gender**: Indicates the gender of the biker. This categorical feature (male/female) was converted to a binary feature, renamed as 'is_male'.
- **is_drunk_biker**: A binary feature indicating whether the biker was intoxicated during the accident.
- **biker_location**: Specifies the area where the bicyclist was positioned at the time of the incident, with values including 'Travel Lane', 'Sidewalk / Crosswalk / Driveway Crossing', 'Non-Roadway', and 'Bike Lane / Paved Shoulder'. This categorical feature was transformed using one-hot encoding.
- **is_drunk_driver**: A binary feature indicating if the driver involved in the accident with the biker was intoxicated.
- **vehicle_type**: Describes the type of vehicle involved in the incident, with values including passenger vehicle, light truck, heavy truck, and two-wheeler. This categorical feature was transformed into one-hot encoding.
- **intersection_type**: Provides context regarding the incident's proximity to intersections or other roadway features, with values 'Intersection', 'Non-intersection', 'Intersection-Related', and 'Non-roadway'. This categorical feature was transformed into one-hot encoding.
- **locality**: Categorizes the area by level of urban development, with values 'Urban (>70% Developed)', 'Rural (<30% Developed)', and 'Mixed (30% to 70% Developed)'. This categorical feature was transformed into one-hot encoding.
- **x**: Represents the longitude of the incident location (continuous feature).
- **y**: Represents the latitude of the incident location (continuous feature).
- **light_condition**: Describes the lighting conditions during the accident, with values including daylight, dark-lighted roadway, and dark-roadway not lighted. This categorical feature was transformed into one-hot encoding.
- **road_surface_type**: Indicates the type of road surface where the incident occurred, with values 'smooth asphalt', 'coarse asphalt', and 'concrete'. This categorical feature was transformed into one-hot encoding.
- **month**: Refers to the month in which the incident occurred, with values corresponding to each month of the year. This categorical feature was transformed into one-hot encoding.
- **weather**: Describes the weather conditions on the day of the incident, with values 'clear', 'cloudy', and 'rain'. This categorical feature was transformed into one-hot encoding.

- **road_condition (T)**: Represents the condition of the road surface at the time of the incident (wet/dry). We transformed this feature into a binary format, renamed as 'is_wet'.

- **severity (Y)**: Indicates the injury severity of the biker in the accident, with values 'no injury', 'possible injury', 'suspected minor injury', 'suspected serious injury', and 'killed'. This feature was transformed into an ordinal scale, with each category mapped to a float from 0 to 1 based on severity.

2.3 Data Filtering

This study focused on incidents occurring exclusively during the winter months—November, December, January, and February—to capture data within North Carolina's wet season, where all recorded bicycle accidents took place. By restricting the analysis to these months, we aimed to maintain a more consistent probability of wet road conditions, reducing potential biases introduced by seasonal variations in weather. Including summer months, which generally have a lower incidence of wet conditions, could introduce skewness and undermine the analysis.

The dataset after preprocessing contained 8,154 observations, with a pronounced imbalance in the road condition variable: 7,522 cases (approximately 92.2%) involved dry conditions, while only 632 cases (around 7.8%) involved wet conditions. Table 1 presents the monthly distribution of road conditions prior to filtering.

Month	January	February	March	April	May	June	July	August	September	October	November	December
Dry (%)	90.3	87.95	92.45	94.96	94.86	94.98	92.28	93.08	6.89	92.62	88.52	83.09
Wet (%)	9.7	12.05	7.55	5.04	5.14	5.02	7.72	6.92	6.89	7.38	11.48	16.91

Table 1: Distribution of road conditions by month

After applying the winter month filter, the dataset was reduced to 1,750 observations, resulting in a more balanced distribution: 1,531 observations for dry conditions and 219 for wet conditions, equating to a wet condition rate of approximately 12.5%. While this adjustment does not fully address the imbalance, it notably improves the representation of wet conditions in the data. Consequently, this refines our research question to: *Does road condition affect the severity of injury in bicycle accidents during the winter?*

2.4 Data Statistics

2.4.1 Distributions

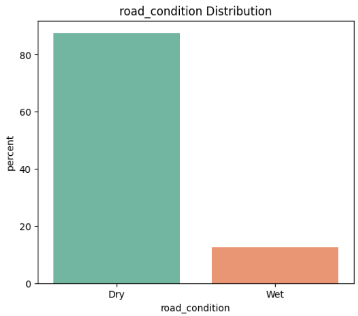


Figure 1: T Distribution

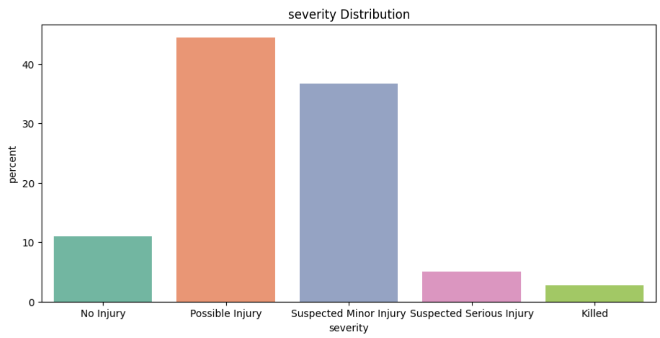


Figure 2: Y Distribution

As can seen the distributions are unbalanced. We will discuss this more detailed in the challenges section.

2.4.2 Chi-Square Test of Independence

We employed the Chi-Square Test of Independence to examine associations between categorical features and both T and Y, aiming to gain statistical intuition regarding potential confounders and effect modifiers. A p-value of less than 0.05 indicates rejection of the null hypothesis H_0, suggesting that the variables may be statistically associated. However, this test only serves to offer preliminary insight.

Feature	Chi Square	p value
is_drunk_biker	40.81	<0.05
biker_location	56.11	<0.05
gender	6.88	0.143
intersection_type	61.92	<0.05
month	15.84	0.199
is_drunk_driver	27.84	<0.05
vehicle_type	28.33	<0.05
light_condition	69.07	<0.05
locality	56.00	<0.05
road_surface_type	7.19	0.516
weather	22.79	<0.05

Table 2: Chi-Square Independence Test Results for Categorical Features with Y

Feature	Chi Square	p value
is_drunk_biker	1.87	0.172
biker_location	4.53	0.209
gender	1.58	0.209
intersection_type	9.81	<0.05
month	10.59	<0.05
is_drunk_driver	0.00	1.0
vehicle_type	2.49	0.477
light_condition	38.94	<0.05
locality	1.80	0.407
road_surface_type	1.76	0.414
weather	1116.79	<0.05

Table 3: Chi-Square Independence Test Results for Categorical Features with T

Based on these results, we identify potential confounders and modifiers:

Potential Confounders: Features that demonstrate significant associations with both T and Y may act as confounders, potentially influencing both variables. In this analysis, 'intersection_type', 'light_condition', and 'weather' emerge as potential confounders, as they exhibit significant relationships with both T and Y.

Potential Modifiers: Effect modifiers are variables that may influence the relationship between T and Y. While these may not show significant associations with both variables, they can affect the impact of one variable on the other. The features 'is_drunk_biker', 'is_drunk_driver', 'locality', 'biker_location', and 'vehicle_type' could be potential modifiers, as they show a significant association with Y but not with T.

2.5 The causal problem definition

Let T be the treatment indicator, where $T = 0$ represents dry road conditions and $T = 1$ represents wet road conditions in the winter. Let Y denote the potential outcomes for injury severity, where $Y \in \{0, 0.25, 0.5, 0.75, 1\}$ corresponds to injury severity levels: no injury (0), possible injury (0.25), suspected minor injury (0.5), suspected serious injury (0.75), and killed (1).

The Average Treatment Effect (ATE) is defined as: $ATE = E[Y_1] - E[Y_0]$.

From this study, we expect to find that wet road conditions ($T = 1$) are associated with higher injury severity (Y) compared to dry road conditions ($T = 0$), indicating that increased wetness leads to more severe injuries in bicycle accidents. Thus, we anticipate that the sign of the ATE will be positive ($ATE > 0$).

2.6 Challenges

- **Unbalanced Distribution**: As discussed in the 2.4.1 section, both the T and Y distributions are imbalanced. This imbalance can introduce bias and affect the results, potentially altering findings if the data were balanced. To reduce the imbalance in the in the T distribution, we applied filtering based on winter months, as described in the 2.3 section.

- **Categorical Features**: The dataset includes numerous categorical features, which were encoded using one-hot encoding. Consequently, the dimensionality of the feature space increased, making it more challenging to model these categorical variables effectively.

- **Missing Data**: A significant portion of records was removed due to missing or unknown values, which may impact the statistical significance of the results.

- **Possible Hidden Confounders**: While the dataset includes a broad range of features, there is always the possibility of unmeasured hidden confounders. One potential hidden confounder in this study is road maintenance quality. While some aspects of maintenance are indirectly represented in the dataset by features such as biker location, intersection type, locality, and road surface type, these do not fully capture the impact of road maintenance. Poorly maintained roads, for instance, may have inadequate drainage, making wet conditions more likely to persist. Simultaneously, these conditions could increase injury severity if an accident occurs, as unmaintained roads often have hazards like potholes or uneven surfaces, which can lead to more severe injuries. Other potential confounders include traffic density and urban planning factors. For example, high-traffic areas may experience distinct road conditions due to frequent use of de-icing agents or salt in winter, which can influence road dryness. Additionally, higher traffic density may lead to more severe collisions, thereby impacting injury severity.

- **Modifiers**: Certain factors in the dataset, such as age, speed limit, and vehicle type, may act as effect modifiers. These variables could influence the relationship between road condition and injury severity by altering the impact of one on the other. For instance, a higher speed limit or larger vehicle may amplify the severity of injuries on wet roads. Similarly, the age of the bicyclist or driver might impact injury outcomes, as younger or older individuals may respond differently to similar crash conditions.

3 Assumptions

- **Stable Unit Treatment Value Assumption - SUTVA**: The SUTVA requires that the treatment or control status of any given unit does not influence the potential outcomes of other units. In the context of this study, each bicycle accident is treated as an independent event, meaning that the treatment condition—whether the road is wet or dry—for one incident has no effect on the severity outcome of other incidents. For instance, wet conditions in one specific location do not affect the crash severity of incidents occurring elsewhere. We made sure that each record in the dataset represents an independent accident, with no interactions between incidents, making SUTVA appropriate for this analysis.

- **Consistency**: The assumption of consistency implies that the observed outcome for any unit is precisely the potential outcome that would occur if the assigned treatment were indeed the treatment given. In this study, consistency guarantees that the observed outcome of injury severity for a bicycle accident under wet conditions is the exact potential outcome expected under wet conditions, and likewise for dry conditions. Given that road conditions are directly observed and recorded at the time of each incident, the consistency assumption is reasonably upheld.

- **Ignorability**: The ignorability assumption asserts that there are no unmeasured confounders that might bias the assignment of treatment. Although it is difficult to verify the complete absence of potential confounders, we contend that this assumption is justifiable in our study due to the comprehensive nature of the data collected. As discussed in Section 2.6, regarding possible hidden confounders, while several may exist, our features partially represent these confounders; therefore, we assume that the influence of any remaining unmeasured confounders is hopefully minimal. In light of these considerations, we maintain that the ignorability assumption is upheld.

- **Common Support**: The common support assumption requires that the probability of observing each treatment level (in this case, wet or dry road conditions) is positive across all covariate profiles. For meaningful comparisons, it is essential that both road conditions are similarly probable across the dataset. In this study, common support is achieved by restricting the dataset to winter-season incidents. By filtering for winter, the probability of encountering wet or dry road conditions is more evenly distributed, ensuring that each sample has an equal likelihood of experiencing either road condition. This avoids the potential imbalance observed in summer, where the probability of wet conditions is considerably lower.

- **Randomness of the Treatment**: The assumption of randomness in treatment assignment implies that treatment (in this case, road condition) should be assigned in a way that approximates randomization for robust causal inference. In the current study, natural variation in road conditions due to seasonal weather patterns serves as an approximation of random assignment.

- **Ordinal Outcome**: We assumed the following order on the outcome (Y) from low to high severity: 'No Injury', 'Possible Injury', 'Suspected Minor Injury', 'Suspected Serious Injury,' and 'Killed'. For that reason, we treated the outcome as an ordinal variable. We believe this approach is reasonable because the injury categories represent progressively increasing levels of harm, and each category implies a higher level of severity than the previous one.

4 Methods

To estimate the ATE, we employ a range of approaches, with a primary focus on propensity score-based methods and learning methods. For each method, we compute the ATE estimator and construct a 95% bootstrap confidence interval (CI) to assess statistical significance.

4.1 Propensity-Based Methods

4.1.1 Propensity Model Evaluation

We evaluated multiple classification models for estimating propensity scores, including Logistic Regression (LR), Gradient Boosting (GB), Random Forest (RF), and XGBoost (XGB).

To select the most appropriate model, each was assessed using train-test split based on the following criteria:

- **Propensity Overlap (Common Support)**: Our goal is to achieve substantial propensity overlap between treatment and control groups. Limited overlap can result in propensity scores that are uninformative and prone to miscalibration.
- **Calibration Curve**: This metric assesses the calibration accuracy of probabilistic predictions made by a binary classifier. An intercept near zero and a slope close to one indicate good calibration across various subgroups or individuals. In contrast, a slope significantly above or below one suggests poor calibration within certain subgroups.
- **Additional Classification Evaluation Metrics**: To comprehensively evaluate each classifier, we used metrics such as Brier score, AUC-ROC, PR-AUC, and log loss. Our aim is to maximize AUC-ROC and PR-AUC values while minimizing Brier score and log loss.

4.1.2 ATE Estimation Methods

We will evaluate the following propensity-based methods for estimating the Average Treatment Effect (ATE):

- **Naive**: This method serves as a naive baseline against which we can compare our results. In this approach, the ATE is estimated as $\overline{Y}_1 - \overline{Y}_0$, where $\overline{Y}_1$ represents the average outcome for the treatment group and $\overline{Y}_0$ represents the average outcome for the control group.
- **IPW**:

$$\widehat{ATE}_{IPW} = \frac{1}{n}\sum_{i=1}^{N}\frac{t_i \cdot y_i}{\hat{e}(x_i)} - \frac{1}{n}\sum_{i=1}^{N}\frac{(1-t_i)y_i}{1-\hat{e}(x_i)},$$

  where $\hat{e}(x_i)$ is estimated by the propensity score model.
- **1NN Matching**:

$$d(x,z) = \sum_{i=1}^{p}|x_i - z_i|,$$

$$CATE(i) = \begin{cases} y_i - y_{j(i)} & \text{if } t_i = 1, \\ y_{j(i)} - y_i & \text{if } t_i = 0, \end{cases}$$

$$\widehat{ATE}_{1NN} = \frac{1}{n}\sum_{i=1}^{n} CATE(i),$$

  where $j(i)$ is the index of the closest neighbor to unit i based on distance d, and $y_{j(i)}$ is the outcome of this closest neighbor.

4.2 Learning-Based Methods

4.2.1 Learning Model Evaluation

We evaluated several classification models to identify the most suitable approach for outcome prediction, specifically examining Logistic Regression (LR), Gradient Boosting (GB), and Random Forest (RF) algorithms. Each model was assessed using a train-test split, with accuracy and F1 scores serving as evaluation metrics, given the multicategorical nature of the outcome. Our objective is to maximize these performance metrics.

4.2.2 ATE Estimation Methods

We will evaluate the ATE based on 2 learning methods:

- **S-learner**:
  - Fit a model with t as a feature on the entire sample: $\hat{y} \approx f(x, t)$.

$$\widehat{ATE}_{S\text{-learner}} = \frac{1}{N} \sum_{i=1}^{N} \left(f(x_i, 1) - f(x_i, 0) \right).$$

- **T-learner**:
  - Fit a model on each treatment value: $\hat{Y}_0 \approx f_0(x)$, $\hat{Y}_1 \approx f_1(x)$.

$$\widehat{ATE}_{T\text{-learner}} = \frac{1}{N} \sum_{i=1}^{N} \left(f_1(x_i) - f_0(x_i) \right).$$

5 Results

5.1 Quality of Classifiers

5.1.1 Propensity-Based Methods

In terms of propensity score overlap in 3, all methods demonstrated a satisfactory overlap between treatment and control groups, with LR and XGB showing the most substantial overlap, providing a substantial common support region and reducing potential bias. Regarding figure 4, LR, GB, and XGB aligned closely with the ideal calibration line, indicating well-calibrated probability estimates. In contrast, RF deviated significantly from the ideal line. On classification metrics in 5, XGB consistently performed well. LR also performed strongly, particularly in PR-AUC, but XGBoost maintained a slight edge overall. In summary, XGB demonstrated the best balance across propensity overlap, calibration, and classification metrics, making it the most suitable model for propensity score estimation in this analysis.

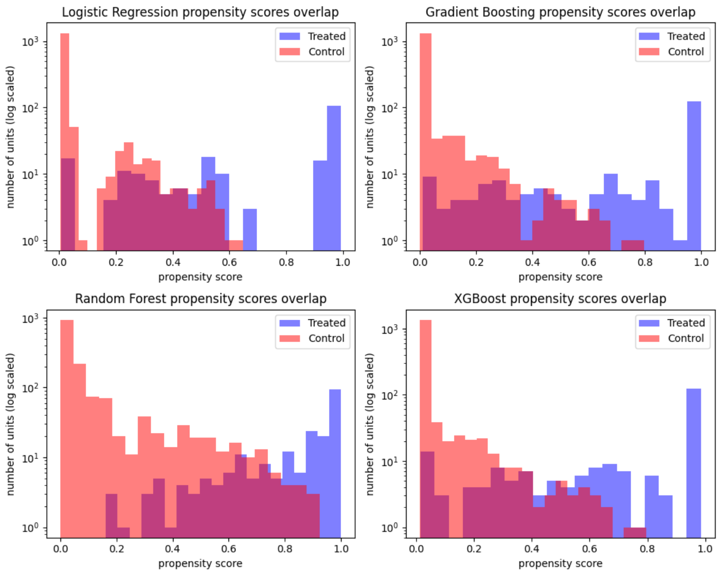


Figure 3: Propensity Score models overlap graph

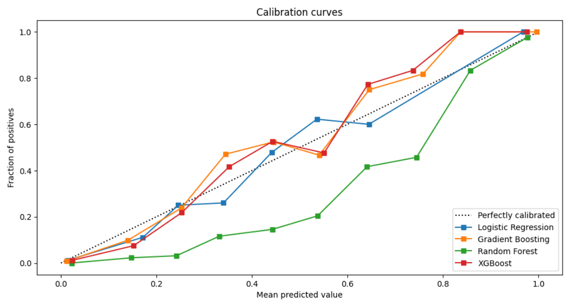


Figure 4: Calibration plots of all propensity models

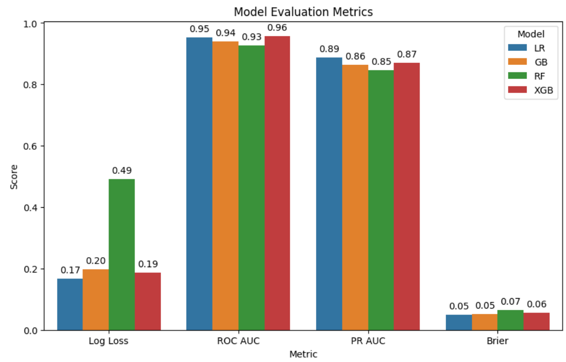


Figure 5: Classification evaluation scores of all propensity models

5.1.2 Learning-Based Methods

Given these results in 6, we chose the RF model because it achieves the highest accuracy and F1 score compared to the other models.

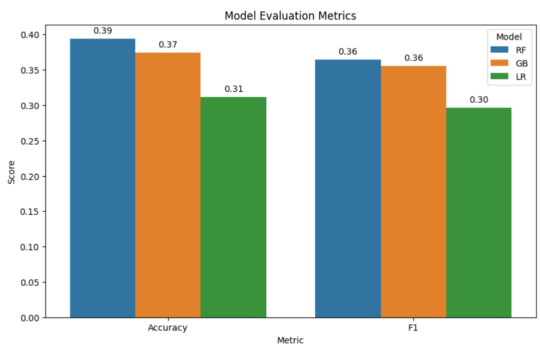
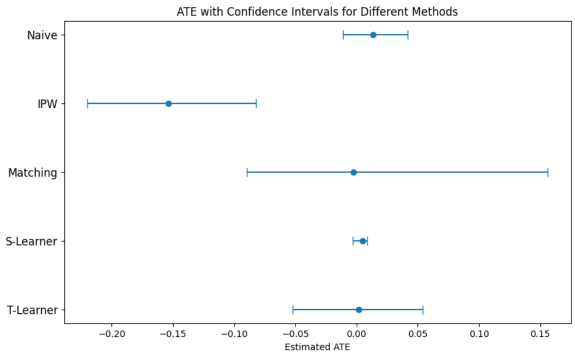


Figure 6: Outcome models evaluation graph

5.2 ATE Estimation

Figure 7a and Table 7b displays the ATE estimates along with their 95% confidence intervals for each method. While the IPW method shows a slightly more negative ATE compared to other methods, the confidence intervals across all methods overlap significantly and include values close to zero. This indicates that, overall, there is no substantial impact of road conditions on injury severity in bicycle accidents based on the current dataset and analysis.



(a)

Method	ATE	Lower Bound	Upper Bound
Naive	0.0136	-0.0114	0.0415
IPW	-0.1536	-0.2200	-0.0820
Matching (1NN)	-0.0028	-0.0896	0.1563
S-Learner	0.0050	-0.0033	0.0089
T-Learner	0.0019	-0.0520	0.0543

(b)

Figure 7: Comparison of ATE Estimates and 95% CIs Across Different Methods

6 Possible Weaknesses

- **Seasonal Limitation**: The dataset was filtered to include only the winter season, as discussed in Section 2.3. Consequently, the results may not be generalizable to other seasons or months, as they may not accurately represent conditions in other seasons, where weather patterns and road conditions differ significantly.
- **Model Dependency**: All the ATE estimation methods evaluated (except the naive approach), are reliant on specific models. The choice of different models or the application of feature selection could potentially lead to varying results.
- **Data Transformations**: During preprocessing, some feature values were aggregated to reduce complexity, as discussed in Section 2.2. This simplification may impact the overall results. While this simplification was intended to decrease noise, it may also lead to information loss.
- **Lack of Data**: After filtering, the dataset was reduced to 1,750 samples, which may limit the reliability of the analysis. A smaller sample size increases susceptibility to outlier effects and may lead to greater uncertainty in ATE estimates, particularly for underrepresented subgroups due to the inherent data imbalance.

7 Discussion

The findings of this study contribute to understanding the impact of road conditions on the severity of injuries sustained in bicycle accidents. Our analysis did not yield a strong causal relationship between wet road conditions and increased injury severity. While previous research

suggested that wet surfaces might increase risks for cyclists as mentioned in 1, our results, which utilized propensity-based and learning-based methods for causal inference, did not indicate a statistically significant ATE. The CIs for the ATE across all methods included values close to zero, suggesting that, within the dataset and seasonal restrictions applied, wet conditions alone may not substantially elevate injury severity.
Several limitations might explain this unexpected outcome. First, while our data pre-processing and filtering helped mitigate imbalance issues, there remains the possibility that the relatively low occurrence of incidents under wet conditions limited statistical power. Additionally, potential hidden confounders, such as road maintenance quality or variations in traffic density, were not directly measurable and might have impacted results. Factors such as traffic levels or road surface maintenance could interact with road wetness in ways that influence accident outcomes, but these effects were not fully captured in our data.
In summary, although wet road conditions have been associated with higher risks in prior studies, our analysis did not find conclusive evidence of a causal link between wet conditions and increased injury severity for cyclists. Further research with more comprehensive data on road maintenance, traffic density, and other factors could provide additional insights.

References

- NSC Injury Facts- Bicycle Deaths
- Asgarzadeh, Morteza, et al (2018). The impact of weather, road surface, time-of-day, and light conditions on severity of bicycle-motor vehicle crash injuries. *American journal of industrial medicine*, 556-565.
- Nyberg, Peter, Ulf Björnstig, and L. O. Bygren (1996). Road characteristics and bicycle accidents. *Scandinavian journal of social medicine*, 293-301.
- Janstrup, Kira Hyldekær, Mette Møller, and Ninette Pilegaard (2018). The road and its influence on bicycle accidents in Denmark. *Transport Research Arena*